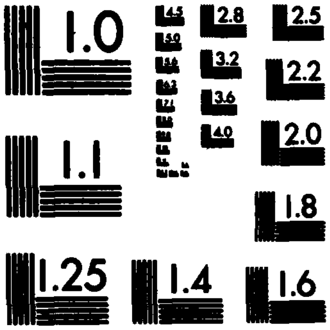
AD-A120 574 HYPOTHESIS TESTING FROM A BAYESIAN PERSPECTIVE(U) 1/1
DECISION RESEARCH EUGENE OR B FISCHHOFF ET AL. JUL 82
PTR-1092-82-6 N00014-80-C-0150

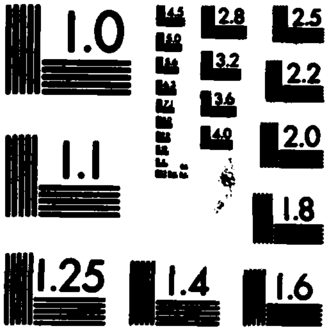
UNCLASSIFIED F/G 12/1 NL

END

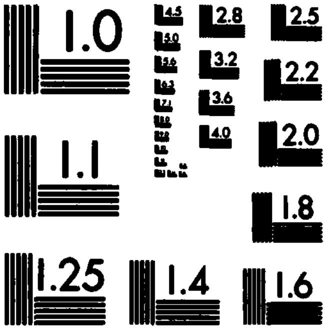
FILMED



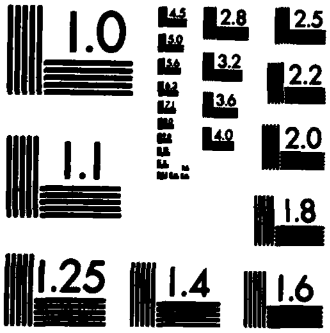
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



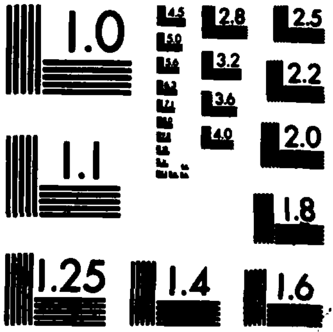
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

12

TECHNICAL REPORT PTR-1092-82-6
CONTRACT NUMBER: N00014-80-C-0150
WORK UNIT NUMBER:NR197-064
July 1982

HYPOTHESIS TESTING FROM A BAYESIAN PERSPECTIVE

Baruch Fishhoff
Ruth Beyth-Marom

Prepared for:

OFFICE OF NAVAL RESEARCH
800 North Quincy Street
Arlington, VA 22217

PERCEPTRONICS

6271 VARIEL AVENUE • WOODLAND HILLS • CALIFORNIA 91367 • PHONE (213) 884-7470

NOTES

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any office of the United States Government.

12

TECHNICAL REPORT PTR-1092-82-6
CONTRACT NUMBER: N00014-80-C-0150
WORK UNIT NUMBER:NR197-064
July 1982

HYPOTHESIS TESTING FROM A BAYESIAN PERSPECTIVE

Baruch Fishhoff
Ruth Beyth-Marom

Prepared for:
OFFICE OF NAVAL RESEARCH
800 North Quincy Street
Arlington, VA 22217

PERCEPTRONICS
6271 VARIEL AVENUE • WOODLAND HILLS • CALIFORNIA 91367 • PHONE (213) 884-7470

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A120 574	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Hypothesis Testing from a Bayesian Perspective		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER PTR-1092-82-6
7. AUTHOR(s) Baruch Fischhoff and Ruth Beyth-Marom		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0150
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Research A Branch of Perceptronics 1201 Oak Street, Eugene, Oregon 97401		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217		12. REPORT DATE July 1982 13. NUMBER OF PAGES 48
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayesian theory
probability assessment
hypothesis testing
conditional probability

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Bayesian inference provides a general framework for testing hypotheses. It is a normative method, in the sense of prescribing how hypotheses should be tested. However, it may also be used descriptively, by characterizing people's actual hypothesis-testing behavior in terms of its consistency with or departures from the model. Such a characterization may facilitate the development of psychological accounts of how that behavior is produced (i.e., as the result of failed attempts to act in a Bayesian fashion, as the result of attempts to process information in non-Bayesian ways. This essay exploits the descriptive potential of Bayesian inference. First, it identifies a set of logically possible forms of non-Bayesian behavior. Second, it reviews existing research in a variety of areas in order to see whether these possibilities are ever realized. The analysis shows that in some situations, several apparently distinct phenomena are usefully viewed as special cases of the same kind of behavior, whereas in other situations previous investigations have conferred a common label (e.g., confirmation bias) to several distinct phenomena. It also calls into question a number of attributions of judgmental bias, suggesting that in some cases the bias is different than what has previously been claimed, whereas in others, there may be no bias at all.

Table of Contents

	Page
Summary	i
Introduction	1
Theory	2
Definition of Probability	2
Updating	3
Likelihood Ratios	3
Action	4
Ways to Stray	5
Hypothesis Formation	6
Assessing Component Probabilities	8
Assessing Prior Odds	11
Assessing the Likelihood Ratio	13
Aggregation	16
Information Search	18
Action	20
Complications	21
Conditional Independence	24
Limits	25
Implications and Reinterpretations	26
Nisbett, Zukier & Lemley (1981)	27
Snyder & Swann (1978)	29
Wason (1960, 1968)	31
Conclusion	34
Reference Note	37
References	38
Acknowledgements	44
Footnotes	45
Distribution List	48

Summary

Bayesian inference provides a general framework for testing hypotheses. It is a normative method, in the sense of prescribing how hypotheses should be tested. However, it may also be used descriptively, by characterizing peole's actual hypothesis-testing behavior in terms of its consistency with or departures from the model. Such a characterization may facilitate the development of psychological accounts of how that behavior is produced (i.e., as the result of failed attempts to act in a Bayesian fashion or as the result of attempts to process information in non-Bayesian ways).

This essay exploits the descriptive potential of Bayesian inference. First, it identifies a set of logically possible forms of non-Bayesian behavior listed according to the task in which the problem arises: hypothesis formulation, assessing component probabilities, assessing prior odds, assessing likelihood ratio, aggregation, information search or action. For example, one potential failure in hypothesis formulation is for the hypothesis to be untestable as a result of being ambiguous and/or complex.

The essay then proceeds to apply this framework in a selective review of existing research in a variety of areas. It does so with a triple purpose: (a) to illustrate the different biases, (b) to identify which of the set of all possible biases have in fact been observed and documented, and (c) to show how the entire hypothesis-testing process must be considered when characterizing any one individual behavior.

Several conclusions emerge from the analysis: (a) In some situations, several phenomena that have previously been thought of as distinct may be usefully viewed as special cases of the same behavior. The neglect of the alternative hypothesis in assessing the likelihood ratio is one such example of a simple phenomenon that has been differently labeled in different contexts (pseudodiagnosticity, inertia, cold readings).

(b) In other situations, previous investigations have conferred a common label to several distinct phenomena (for example, "confirmation bias" has been applied to acts that might be better characterized as "failure to ask potentially falsifying questions", "asking non-diagnostic questions", "mistaking affirmation for confirmation", etc.).

(c) It calls into question a number of attributions of judgmental bias, suggesting that in some cases the bias is different than what has previously been claimed, whereas in others, there may be no bias at all.

Hypothesis Testing from a Bayesian Perspective

Hypothesis testing is a crucial intellectual activity. Not surprisingly, it is also a focus of psychological research. A variety of methods have been applied to understand how people gather and interpret information in order to evaluate hypotheses. Either implicitly or explicitly, some theory of how people should test hypotheses provides the conceptual framework for these studies of how they do test them. Such a theory provides a set of articulated terms for describing tasks and a definition of appropriate behavior against which actual performance may be compared. The theory may even be psychologically valid, at a certain level, if people are found to follow its dictates, either due to natural predilections or because they have been trained to do so. Even when behavior is sub-optimal, some psychological insight may be obtained by asking whether that behavior may be described as some sort of systematic deviation from the theory. At the very least, reference to a normative theory can identify performance deficits that need to be understood and rectified.

One general and popular normative scheme is Bayesian inference, a set of procedures based upon Bayes' Theorem. These procedures show how to (a) identify the data sources that will be most useful for discriminating between competing hypotheses, (b) assess the implications of observed data vis-a-vis the truth of competing hypotheses, (c) aggregate the implications of different data into an overall appraisal of hypothesis validity, and (d) use that appraisal to select the course of action that seems best in the light of available evidence. Excellent detailed expositions of the scheme may be found in Edwards, Lindman and Savage (1963), Phillips (1972), and Schlaifer (1969).

The present essay begins by presenting a rudimentary version of Bayesian inference. From this simple scheme, it derives a taxonomy of logically possible deviations. This taxonomy is then used to characterize a variety of published studies reporting biased hypothesis testing. In some cases, the result is to reiterate the claims of the original investigators; in other cases, those claims are countered by alternative interpretations, suggesting other biases that may be involved or ways in which observed behavior might be construed as being properly Bayesian; in still other cases, research conducted in other traditions is cast in Bayesian terms, in the hopes of profiting from others' experience and drawing different fields together.

Rather than being exhaustive, this literature review is meant to be illustrative of how the Bayesian perspective can be used to illuminate various tasks. As a result, it will emphasize new interpretations and reinterpretations over repetition. Similarly, it will focus on those aspects of hypothesis testing that have been given the least psychological treatment, namely, the recruitment and evaluation of evidence. By reference to the taxonomy, it will also identify potential biases for which positive evidence is lacking--prompting the question of whether it was an opportunity for research or an opportunity for error that has been missed.

Theory

Definition of Probability

From the Bayesian perspective, knowledge is represented in terms of statements or hypotheses, H_i, each of which is characterized by a subjective probability, $P(H_i)$, representing one's degree of belief in its truth. For example, one might be .75 confident that "it will snow tomorrow." Such probabilities are entirely subjective, in the sense that different individuals may legitimately assess quite different probabilities. The term "assess" is used rather than "estimate" in order to emphasize the notion that in providing a probability one is attempting to express one's own feelings rather than trying to appraise a property of the physical world. Thus, there is no "right" probability value for a particular statement. Even if a very low probability proves to be associated with a true statement, one cannot be sure that it was not an accurate reflection of the assessor's (apparently erroneous) store of knowledge.

The constraints on subjective probabilities emerge when one considers sets of assessments. Formally speaking, a set of probabilities should be orderly or coherent, in the sense of following the probability axioms (Kyburg & Smokler, 1980).[1] For example, $P(H) + P(\bar{H})$ should total 1.0.

Updating

Additional rules derived from the probability axioms lead to Bayes' Theorem, which governs the way in which one's beliefs in hypotheses should be updated in the light of new information. In its simplest form, the theorem treats the implications of an observation that produces the datum D for determining whether an hypothesis H is true, relative to its complement, $\bar{H}$. In such cases, Bayes' Theorem states that

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(D|H)}{P(D|\bar{H})} \cdot \frac{P(H)}{P(\bar{H})} \qquad (1)$$

Reading from the right, the three terms in this formula are: (a) prior odds that H (and not $\bar{H}$) is true in the light of all that is known before the receipt of D; (b) the likelihood ratio, representing the information value of datum D with respect to the truth of H; (c) the posterior odds that H is true in the light of all that is known after the receipt of D.

Likelihood Ratios

If the probability of observing D given that H is true is different from the probability of observing D when H is not true, then the likelihood ratio is different from 1 and the posterior odds are different from the prior odds. That is, the odds favoring H are smaller or greater as a result of having observed D than they were before. Such a datum is considered to be informative or diagnostic. Its diagnosticity can be expressed as the magnitude of the likelihood ratio or its reciprocal, whichever is larger. Clearly, diagnosticity depends upon the hypotheses being tested. A datum that distinguishes one hypothesis from its complement may be completely uninformative about another pair of hypotheses. Data do not answer all questions with equal efficacy.

The value of the likelihood ratio is independent of the value of the prior odds. One could, in principle, observe a datum strongly supporting an hypothesis that is initially very unlikely (although, of course, one would not expect to). Conversely, one's posterior odds favoring H might be very low after observing a datum that supports H if one's prior odds had been even lower.

There is also no necessary relationship between the values of the numerator and the denominator of the likelihood ratio. A datum might strongly favor H even if it is very unlikely given the truth of H, provided it is even more unlikely given the truth of $\bar{H}$. Similarly, observing a datum that is a necessary concomitant of H (i.e., $P(D|H \simeq 1)$) may be uninformative if it is also a necessary concomitant of $\bar{H}$.

Action

The apparatus of Bayesian inference also provides tools for converting one's beliefs in hypotheses into guides to action. In simplest terms, these tools translate the cost associated with erroneously acting as though an hypothesis is true and the cost of erroneously acting as though an hypothesis is false into a critical ratio. If the posterior odds favoring H are above this value, then one is best off acting as though H is true; if they are below, then one is best off acting as though H is false.

The value of the critical ratio will, of course, depend on the particular kinds of action that are contemplated. Where one's posterior odds stand, vis-a-vis the critical ratio, will depend upon both what one's prior odds were and what evidence one has subsequently received. Two individuals who agreed on the costs of the different errors and on the meaning of the different data might still act differently, if they had different prior odds. On the other hand, if the cumulative weight of the new evidence was sufficient, they might act in the same way despite quite discrepant beliefs (e.g., if the evidence carried them from priors of 1:10 and 1:1 to posteriors of 100:1 and 1000:1, respectively).

When one has the option of collecting more data, additional Bayesian procedures can help one select the most useful observation. Such value-of-information analyses evaluate the expected impact of each observation on the expected utility of the actions one can take. They can also tell when the cost of further observations is greater than their expected benefit.

Ways to Stray

Bayesian inference, like other normative schemes, regards only those who adhere to its every detail as behaving optimally. Conversely, every component of the scheme offers some opportunity for error. The present section offers a catalogue of possible pitfalls. For the normatively minded, these possibilities might be seen as defining Bayesian inference negatively, by pointing to behavior that is inconsistent with it.

For the descriptively minded these logical possibilities suggest judgmental biases that might be observed in empirical studies. If observed in situations in which people are properly instructed and motivated to respond correctly, such deviations can be theoretically informative because they would seem to reflect deep-seated judgmental processes. From a practical standpoint, such deviations suggest opportunities for constructive interventions that might lead to better inferences and subsequently to better decisions based on those inferences. These interventions might include training, the use of decision aids, or the replacement of intuition by formalized procedures (Edwards, 1968; Fischhoff, 1982; Kahneman & Tversky, 1979).

These potential biases are presented telegraphically in Table 1. The left-hand column shows the task (or subtask) in which the problem arises; the center provides the possible biases; the right-hand column points to phenomena reported in the literature that we have interpreted as special cases of these biases. These are described in detail below.

Table 1

Potential Sources of Bias in Bayesian Hypothesis Testing

Task	Potential Biases	Special Cases
Hypothesis formation	untestable non-partition	ambiguity, complexity, evidence unobservable non-exclusive, non-exhaustive
Assessing component probabilities	misrepresentation incoherence miscalibration non-conformity objectivism	strategic responses; non-proper scoring rules non-complementarity; disorganized knowledge overconfidence reliance on availability or representativeness ---
Assessing prior odds	poor survey of background failure to assess	incomplete, selective base-rate fallacy
Assessing likelihood ratio	failure to assess distortion by prior beliefs neglect of alternative hypotheses	non-causal, "knew-it-all-along" preconceptions, lack of convergence pseudodiagnosticity, inertia, cold readings
Aggregation	wrong rule misapplying right rule	averaging, conservatism? computational error, conservatism?
Information search	failure to search non-diagnostic questions inefficient search unrepresentative sampling	premature conviction tradition, habit failure to ask potentially falsifying questions ---
Action	incomplete analysis forgetting critical value	neglecting consequences, unstable values confusing actual and effective certitude

Hypothesis Formation

For hypothesis testing to begin, there must be hypotheses to test. Indeed, because the diagnostic impact of data is defined only in the context of particular hypotheses, there is no systematic way that data can even be collected without such a context. In the absence of any hypotheses, the collection of data would represent idle stockpiling. Although a logical possibility, this would not seem to be a troublesome or common bias. In practice, even the most ambling data collection may be guided by at least a vague idea of the kinds of hypotheses that the collector might someday be asked to test.

A more serious threat than the absence of hypotheses to test is the possession of untestable ones. One route to untestability lies through ambiguity, either intentional or inadvertent. To take a popular example, astrology columns offer hypotheses about what consequences will follow what acts (e.g., "You will be better off avoiding risky enterprises"). Yet these acts and consequences are so vaguely defined that it is unclear whether what actually happens affirms or disconfirms the hypothesis. By way of contrast, probabilistic risk analysts generate highly detailed hypotheses about the operation of technical systems (Green & Bourne, 1972; U.S. Nuclear Regulatory Commission, 1981; e.g., "toxins can be released to the atmosphere only if the following events occur..."). Although actual operating experience should afford an opportunity to test these hypotheses, it may be unclear whether the specific events that were observed are subsumed in the generic events described in the analysis. For example, investigators disagreed over whether the fire at the Browns Ferry nuclear power plant in 1975 was included among the accident sequences described in the then-definitive analysis of reactor operation (U.S. Nuclear Regulatory Commission, 1975; 1978).

Complexity offers a second route to untestability. Hypotheses that are set out clearly may have such great detail that no datum provides a clean test. For example, political advisors may escape charges of having predicted events incorrectly by noting that every last detail of their advice was not followed ("Had they only listened to me and done X and Y as well, then everything would have been all right").[2] O'Leary, Coplin, Shapiro, and Dean (1974) found that among practitioners of international relations (i.e., those working for government services) theories are of "such complexity that no single quantitative work could even begin to test their validity" (p. 228). Indeed, some historians argue that the accounts of events that they produce are not hypotheses at all, but attemptr to integrate available knowledge into a coherent whole. In this light, a valid explanation accommodates all facts, leaving none to test it. This attitude toward hypotheses has its own strengths and weaknesses (Fischhoff, 1980).

A third route to untestability is to generate hypotheses for which the relevant evidence is unobservable. For example, a prominent hypothesis in political discussions is that possession of strategic weapons reduces the probability of war (Freedman, 1981). Unless that probability can be measured, this is not an empirical question. The observance or non-observance of war provides no direct evidence of its probability (except that it was not 0 or 1). These debates could be scientific only if some surrogate was found for the probability of war, such as measures of closeness to war or the intentions of national leaders (Stech, 1980)--in which case the hypothesis itself would be somewhat different.

These problems may afflict any effort at hypothesis testing. The Bayesian scheme imposes additional constraints. If it is to produce posterior probabilities, then the hypotheses not only must be testable, but also must partition some space of possibilities. This requirement is fulfilled when one considers an hypothesis and its complement, as well as when that complement is decomposed into a set of mutually exclusive and exhaustive possibilities.[3] Computationally, the scheme does not work with nonexclusive hypotheses, which naturally render the message of evidence ambiguous (since it can support or contradict both).

Whereas achieving mutually exclusive hypotheses requires precision of formulation, securing an exhaustive set of hypotheses often requires the exercise of imagination. Unless there is a natural partition of hypotheses (e.g., H and $\bar{H}$; animal, vegetable and mineral), one must think of all the possibilities. The difficulties of exhaustive enumeration are often exploited by mystery writers, who unearth a neglected hypothesis that tidily accounts for available evidence. For some practical purposes, it is possible (or necessary) to list major alternative hypotheses and group the remainder under "all other possibilities." The success of this strategy depends upon one's ability to estimate the completeness of the set of enumerated hypotheses. Evidence suggests that people tend to exaggerate the completeness of hypothesis sets (Fischhoff, Slovic & Lichtenstein, 1978; Mehle, Gettys, Manning, Baca & Fisher, 1981).

Assessing Component Probabilities

Even the minimal exercise in Bayesian inferences represented in (1) requires one to assess four subjective probabilities, expressing one's degree of belief in the truth of the statements, H, $\bar{H}$, $D|H$, and $D|\bar{H}$. Any difficulties in assessing these component probabilities would impair effective hypothesis testing. As mentioned, the Bayesian perspective holds probabilities to be subjective expressions of belief, reflecting what the assessor knows about the question. Hence, there are no right (or objective) probabilities. Accepting the subjectivist position does not, however, mean accepting any probability that someone produces as being an appropriate assessment of his or her state of belief. There are a number of ways in which the assessment of component probabilities can go wrong.

One possible problem is lack of candor. People may misstate their beliefs, perhaps to give a response that is expected of them, perhaps to avoid admitting that some unpleasant event is likely, perhaps to achieve some strategic advantage by misrepresenting how much they know or what they believe. In situations where the truth of the statements can, at some point in time, be ascertained, candor may be encouraged by the use of proper scoring rules (Murphy, 1972). These rules reward people as a function of both their stated beliefs and the (eventually revealed) state of the world in such a way that the probability with the highest expected value is the one expressing one's true belief. Whether these rules prove effective in practice is a moot point (Lichtenstein, Fischhoff & Phillips, 1982). Where they do not, or where they cannot even be applied (because the truth will not be known or because no reward system is possible), other means of assuring candor are needed.

A second possible problem is doing a poor job of assessing one's knowleldge. One possible symptom of having not put together all that one knows is _incoherence_, failure to follow the probability axioms. For example, $P(H) + P(\bar{H})$ must equal 1.0. Although seemingly trivial, this requirement may be particularly prone to violation when the two probabilities are not assessed simultaneously--and concentration of each hypothesis evokes a somewhat different subset of one's knowledge.

A second possible symptom of poor assessment is _miscalibration_, failure of one's confidence to correspond to reality. If an hypothesis that was initially assigned a probability of being true of .2 was eventually found to be true (beyond any reasonable doubt), the initial assessment may seem somewhat dubious. However, it may have been an accurate summary of the assessor's knowledge at that moment. Recurrent association of high probabilities with false hypotheses would suggest something drastically wrong with the assessor's assessment procedure. Calibration offers a formalization of this reliance on the eventually accepted validity of hypotheses in order to validate prior probability assessment. For the well-calibrated assessor, probability judgments of .XX are associated with value hypotheses XX% of the time. Empirical studies of calibration have found that probabilities are related, but not identical, to proportions of correct hypotheses. The most common deviation is overconfidence; for example, being only 80% correct when 1.00 confident (Lichtenstein, Fischhoff & Phillips, 1982).

A third symptom is non-conformity, producing a probability that differs from that of "expert assessors" for no apparent defensible reason. The existence of such consensus is most likely when there is a statistical data base upon which to base probability assessments (e.g., public health records of mortality). In this restricted realm, the distinction between subjective and objective probabilities becomes blurred, as subjectivists would typically concur with the relative frequency interpretation of probability which objectivists consider to be the only meaningful one.

The difficulty with relying on incoherence, miscalibration, or non-conformity to signal poor assessment processes is that all three could simply reflect problems with the assessors' knowledge base. Their probabilities may be incoherent because their beliefs are incoherent; the bits and pieces of knowledge that they have about a topic may not have coalesced into an integrated view capable of generating consistent assessments (Lindley, Tversky & Brown, 1979). Similarly, miscalibration or non-conformity may reflect "systematic" misinformation; for some reason or other, what one knows about a topic leads to inapproprite degrees of belief.

A surer way to fault the assessment process itself is to demonstrate that it has followed procedures that are inconsistent with the rules of statistical inference. Two well-known representatives of this genre of argument are the claims that people rely upon the availability and representativeness heuristics when making probability assessments (Kahneman, Slovic & Tversky, 1982; Tversky & Kahneman, 1974). Users of the former judge an event to be likely to the extent that exemplars are easy to recall or imagine; users of the latter judge an event to be likely to the extent that it represents the salient features of the process that might produce it. Although both rules can provide good service, they can also lead the user astray in predictable ways. For example, reliance on availability will induce overestimation of unusually salient events (e.g., the probability of dying from flashy, hence overreported, causes of death such as tornadoes and homicide; Lichtenstein, Slovic, Fischhoff, Layman & Combs, 1978).

A final process problem would be refusal to consider anything but relative frequency data when assessing probabilities. Although subjectivists acknowledge the potential relevance of such data, they will not be bound to it. Indeed, among the selling points of Bayesian inference is that it can accommodate not only diverse kinds of data in the course of a particular problem, but also that each datum may be any conceivable kind of information. One can, for example, assess probabilities for a meaningful shrug or an off-hand comment as well as for a bead drawn from an urn or a 40-subject experiment. The only difference is that as one moves from beads to shrugs, it becomes increasingly difficult to attest to the (un)reasonableness of a particular response. An assessor who failed to seek useful and available non-frequentistic evidence would, from a Bayesian perspective, be foolish. An individual who ignored non-frequentistic evidence that was already on hand would be biased.[4,5]

Assessing Prior Odds

Prior odds represent the ratio of the probabilities of two hypotheses prior to the collection of further evidence. As a result, the difficulties in assessing prior odds are by and large the sum of the difficulties of formulating hypotheses and assessing component probabilities--as discussed above. There seem, however, to be at least two additional (and incompatible) biases that may affect this particular stage of the assessment process.

One such bias is not treating the two component probabilities equally. An extreme form of unfairness is to neglect one of the hypotheses. When people act as though an hypothesis that is most probably true is absolutely true, they have, in effect neglected its complement. In that case, hypothesis testing never begins, because the hypothesis is treated as fact. Even when both hypotheses are considered, one may be given deferential treatment. Skeptics, for example, may give undue weight to evidence that contradicts a favored hypothesis; they have a warm spot in their hearts for complements. On the other hand, Koriat, Lichtenstein, and Fischhoff (1980) found favoritism for initially favored hypotheses when they asked subjects to determine the relative likelihood of two possible answers to a question. Subjects seemed to ask themselves, "Do I know of any supporting evidence?"[6] Such directed search may serve legitimate purposes (e.g., seeing if any case at all can be made for, or against, a particular hypothesis). However, it may be very difficult to estimate (and correct for) the bias that it induces in the resultant sample of evidence. Indeed, it is the failure to realize the unrepresentativeness of the samples that the availability heuristic produces which makes it a potentially dangerous judgment rule.

These difficulties vanish in the presence of another bias that has attracted considerable attention of late: simply neglecting the base rate (Kahneman & Tversky, 1973). The "base-rate fallacy" refers to the tendency to allow one's posterior odds to be dominated by the information that one extracts from the additional datum, D, in neglect of prior odds expressing what typically has been observed. The several recent reviews of this literature (Bar-Hillel, 1980; Bar-Hillel & Fischhoff, 1981; Borgida & Brekke, 1981) may be summarized roughly as indicating that people rely most heavily on whatever information seems most directly pertinent to the task at hand. Thus, for example, when testing a pair of hypotheses such as "John is a lawyer/John is an engineer," even weakly diagnostic information relating directly to John may dominate base-rate information reporting the overall prevalence of the two groups. Base-rate information will, however, be used if it can be linked more directly to the inference or if the case-specific information appears palpably worthless.

A particularly disturbing possibility is that this pattern of results reflects a general aversion to combining different kinds of evidence. If such an aversion reflected a principled reluctance to combine, then it would mean that people are intuitively opposed to Bayesian inference, which is predicated on combination. If such aversion was due to people not knowing how to effect combinations, then it would support the need for Bayesian inference, which provides a general purpose integration rule.

The evidence for the base-rate fallacy comes primarily from studies in which base-rate information was presented, yet ignored. As such, it could be viewed as a problem of aggregation (and hence be treated two sections below). It is discussed here because the failure to use explicitly presented base rates strongly suggests that they will not be spontaneously sought or assessed. This suspicion is confirmed, among other places, by Lyon and Slovic's (1976) finding that, when asked directly, half of their subjects did not believe that base rates were relevant to their judgments.

Assessing the Likelihood Ratio

In principle, people could ignore the likelihod ratio just as well as the base rate, thus one could speak of the "likelihood-ratio fallacy" whenever people fail even to compute the likelihood ratio associated with a pertinent datum. This may happen, for example, when the datum provides merely circumstantial evidence, when it cannot be woven into a causal account involving the hypothesis (Tversky & Kahneman, 1980), or when it reports a non-occurrence. A classic example of the latter is Sherlock Holmes' observation (Doyle, 1974) that his colleague, Inspector Gregory, had not considered the significance of a dog failing to bark (when an intruder approached, indicating their familiarity with one another).

Failure to assess the likelihood ratio of received evidence may also be encouraged by hindsight bias, which leads people to underestimate the informativeness of new data (Fischhoff, 1975; 1982). The feeling that one "knew all along" that D was true might make the calculation of D's likelihood ratio seem like a pointless exercise. Denying that D has anything to add does not, however, mean that it will not have any impact--only that one will be unaware of what that impact has been. At the same time as people deny its contribution, new information may change their thinking in ways that they cannot appreciate or undo (Fischhoff, 1977; Sue, Smith & Caldwell, 1973). These unintended influences may or may not be those that would follow more purposeful deliberations.

When judges do choose to assess the likelihood ratio, the sequential position of that operation may expose it to the influence of the operations that precede it. In particular, the interpretation of new evidence may be affected by previous beliefs, thereby subverting the independence of the likelihood ratio and prior odds. Nisbett and Ross (1980) offer an impressive catalogue of ways in which people can interpret and reinterpret new information so as to render it consistent with their prior beliefs. So great is people's ability to exploit the ambiguities in evidence to the advantage of their preconceptions and to discount the relevance (or credibility) of inconsistent evidence, that erroneous belief can perseverate long after they should have been abandoned.[7]

Such biased interpretation of evidence will also thwart the effective convergence of belief that should follow use of Bayesian inference. However discrepant the initial position of two individuals, their posterior odds should converge for practical purposes, providing they observe a sufficiently large set of diagnostic data about whose interpretation they agree. The ability to interpret the same datum as supporting contradictory hypotheses means that convergence may never occur. Whatever data the two observers see, they become more convinced of their respective rectitude.

When people choose to evaluate evidence, they must compute two conditional probabilities, $P(D|H)$ and $P(D|\bar{H})$. Both assessments are necessary because there is no necessary relationship between these two components of the likelihood ratio. A variety of studies suggest, however, that people consider only the numerator. That is, they are interested in how consistent the evidence is with the hypothesis they are testing, $P(D|H)$, failing to consider its consistency with the alternative hypothesis, $P(D|\bar{H})$. As a result, the size of $P(D|H)$ determines D's support for H. Users of this strategy act as though they assume that the two conditional probabilities are inversely related, although, in principle, both may be high or low. A datum with a low $P(D|H)$ may provide strong evidence for H if $P(D|\bar{H})$ is even lower; a datum for which $P(D|H)$ is high may reveal nothing if $P(D|\bar{H})$ is equal.

Four examples will give some flavor of the variety of guises within which incomplete appraisal of the likelihood ratio may emerge:

(a) Doherty, Mynatt, Tweney, and Schiavo (1979) presented subjects with six pieces of data, D_i; and allowed them to inquire about the values of 6 of the 12 conditional probabilities: $P(D_1|H)$, $P(D_1|\bar{H})$, ... , $P(D_6|H)$, $P(D_6|\bar{H})$. They found that very few subjects requested any of the pairs of conditional probabilities (e.g., $P(D_3|H)$ and $P(D_3|\bar{H})$) needed to compute likelihood ratios. The authors labeled this tendency to pick but one member of each pair, pseudodiagnosticity.

(b) Troutman and Shanteau (1977) had subjects draw beads from a box whose composition was either (70 red, 30 white, and 50 blue) or (30 red, 70 white, and 50 blue) in order to infer whether the box was predominantly red or white. The draw of a blue bead reduced subjects' confidence in whichever hypothesis they were leaning toward, even though a blue bead is equally unlikely under either hypothesis. Thus, they tested only their favored hypothesis and did that only by reference to $P(D|H)$.

(c) In a similar experiment (without the blue balls), Pitz, Downing and Reinhold (1967) found that subjects who were fairly confident that the box was, say, predominantly red would increase their confidence in that hypothesis after observing a white. Subjects apparently felt that they should see an occasional white and apparently neglected to consider that that event was still more likely given that a predominantly white box was being used.

Pitz et al. called this failure of inconsistent evidence to slow the increase of confidence in H, an inertia effect.[8]

(d) One favorite ploy of the magicians, mentalists, and pseudopsychics who claim to read other people's minds is to provide universally valid personality descriptions (Forer, 1949; Hyman, 1977), which apply to almost everyone (although this is not transparently so). They can trust their clients to assess P(this description|my mind is being read) and not P(this description|my mind is not being read).

A final threat to the validity of assessed likelihood ratios comes from the fact that the probabilities involved all concern conditional events. This added complexity seems to make probability assessment rather more difficult, with people forgetting the conditioning event, reversing the roles of the two events, or just feeling confused (Eddy, 1982; Moskowitz & Sarin, Note 1).

Aggregation

Assuming that judges have attended to and assessed all components of the Bayesian model, they must still combine them in order to arrive at posterior odds. The two logically possible sources of error here are using the wrong aggregation rule (e.g., averaging, rather than multiplying, the likelihood ratio and prior odds) and using the right rule, but applying it inappropriately (e.g., making a computational error). Establishing whether either or both of these biases occur in practice was a focus of early research into intuitive Bayesian inference (excellently reviewed by Slovic & Lichtenstein, 1971).

Most of these early studies used the rather un-Bayesian strategy of creating inferential tasks in which the experimenter could claim to know the correct subjective probability for all participants. This was done by using highly artificial stimuli for which one could argue that all reasonable observers would have the same subjective probability. For example, subjects might be shown a series of poker chips and be asked to evaluate the hypotheses: they are being drawn from a bookbag with 70%

blue chips and 30% red chips; they are being drawn from a bag with 70% red chips and 30% blue. The predominant result of this research was that subjects' confidence in the apparently correct hypothesis did not increase as quickly as the accumulating evidence indicated that it should.

A likely debate ensued over whether this poor performance (called conservatism) reflected failure to appreciate how diagnostic the evidence was (called misperception) or failure to combine those diagnosticity assessments according to Bayes' rule (misaggregation). Aside from its theoretical interest, this dispute had considerable practical importance. If judges could assess component probabilities correctly, but could not combine them, they they could be integrated into person-machine systems wherein the machine relieved them of the mechanical computations that they found difficult. (Moreover, it would incorporate an elicitation scheme that kept users from ever forgetting any component probabilities.) On the other hand, if people are the only source of probabilities that they cannot assess very well, then the machine may be just spinning its wheels (Edwards, 1962).

Although the source of this conservatism was never resolved, the sort of hypotheses that were raised reflects the sort of problems that could pose barriers to proper aggregation. Some examples (details of which may be found in Slovic & Lichtenstein, 1971): (a) anchoring: people tend to stay stuck to their own previous estimates; (b) response bias: reluctance to give extreme responses pushes people's answers toward the center of the response range; (c) ceiling effect: fear of "using up" the probability scale makes people hedge their responses; (d) nonlinearity: a given piece of evidence should be associated with the same ratio of posterior odds to prior odds whenever it is received; people may try instead to make the differences between prior and posterior probabilities of H constant; (e) response mode: use of odds (instead of probabilities) as a response mode produces more optimal responses, suggesting that it is a more natural way for people to think about problems and to express their knowledge.

In the end, this line of research was quietly abandoned without establishing the relative roles of these different factors. This cessation of activity seems due to the discovery of the base-rate fallacy (which represents the antithesis of conservatism) and other phenomena that lead many researchers to conclusions such as: "it may not be unreasonable to assume that ... the probability estimation task is too unfamiliar and complex to be meaningful" (Pitz, Downing & Reinhold, 1967, p. 392), "evidence to date seems to indicate that subjects are processing information in ways fundamentally different from Bayesian ... models" (Slovic & Lichtenstein, 1971, p. 728), or "in his evaluation of evidence, man is apparently not a conservative Bayesian; he is not Bayesian at all" (Kahneman & Tversky, 1972, p. 450).

Information Search

Obviously, Bayesian updating requires the collection of additional information beyond what was incorporated in the prior odds. Whether collection is contemplated at all should depend upon one's a priori confidence in the truth of the hypothesis, and whether that is adequate for whatever actions are being considered (a topic considered in the following section). When one would like to know more, the data that one collects (if any) should depend upon the opportunities presented.

These opportunities may be conceptualized as questions that can elicit a set of possible answers (or D_i), each of which carries a message regarding the truth of the hypotheses. All other things being equal (e.g., the cost of asking), the most valuable questions are those that are expected to produce the most diagnostic answers. Conversely, one should never ask questions all of whose possible answers have likelihood ratios of 1. Such questions cannot change one's beliefs regarding the truth of the hypotheses, nor should they affect the choice of actions based on those beliefs. Value-of-information analysis includes a variety of schemes for deciding how much one should spend for information in general and how to devise the most efficient sampling strategies. It considers such factors as the cost of asking, the consequences of the possible decisions, the a priori probability of receiving the different possible answers, and the likelihood ratios associated with those answers (Brown, Kahr & Peterson, 1976; Raiffa, 1968).

Why might someone disregard these considerations and ask questions whose answers cannot have a diagnostic impact? Tradition is one possible reason. One may always have collected a particular datum, without having seriously analyzed what one has learned from it. Official forms and graduate school applications might be two familiar homes for non-diagnostic questions. These traditions may spawn or be spawned by beliefs about the kinds of evidence that are inherently more valuable. In various circles, secret, quantitative, or introspective information might have this special status of always meriting inquiry (Fischer, 1970). Misdirected search may occur also when people's task changes and they fail to realize that the questions that helped to answer the old hypotheses are no longer as effective in answering the new ones. For example, a psychiatric social worker who moved from private practice to a large public agency might require a whole new set of question-asking skills.

The factors leading to pointless questions can, in less extreme form, lead to inefficient ones, questions that provide some information, yet less than the maximum possible. Search problems may be aggravated by difficulties with other aspects of the inferential process. Value-of-information analysis built around preposterior analysis, anticipating what one will believe (and do) if various possible answers are observed. If people cannot assess likelihood ratios properly, then they cannot appraise properly the expected value of possible questions.

A noteworthy form of inefficiency is ignoring the opportunity to ask potentially falsifying questions, ones whose answers have a reasonable chance of effectively ruling out some hypothesis (e.g., a physician who failed to order a test that could eliminate the possibility of a particular disease). What constitutes a "reasonable chance" depends upon the usual value-of-information analysis factors (e.g., the prior probability of the disease, the importance of its detection, the cost of the test).

An obvious threat to sampling is inadvertently selecting an unrepresentative set of information, hence arriving at erroneous beliefs. When a proper sampling frame is available (e.g., for the Census), one can describe a variety of specific violations of representative sampling (Kish, 1965), which people may or may not intuitively appreciate (e.g., Bar-Hillel, 1979). Such situations may, however, be relatively rare with Bayesian inference, where information can come from a variety of sources and in a variety of forms. Good sense then becomes the only guide to sampling. As the history of scientific progress shows, it may take a fortuitous (if unpleasant) surprise to reveal an unintended bias in sampled information.[9]

Action

Although it is possible to undertake hypothesis testing for its own sake, Bayesian inference is also embedded in statistical decision theory. Its output, posterior odds, provides a summary of one's beliefs that facilitates selecting the optimal course of action. Conversely, it is only knowledge of the possible actions and their associated consequences that allow one to determine what information it is best to sample. Two Bayesian judges who considered different actions (or evaluated their consequences differently) might justifiably formulate different hypotheses and sample different data even though they function in the same information environment.

Many of the difficulties that frustrate attempts to take prudent action on the basis of available knowledge are not unique to Bayesian inference, and need not be treated in detail here. These include not having well-articulated values (i.e., not knowing what one wants), failing to think through all the consequences of different actions, and allowing one's preferences to be manipulated by the way in which problems are presented (Fischhoff, Slovic & Lichtenstein, 1980; Tversky & Kahneman, 1981).

With other familiar problems, the Bayesian framework may offer an illuminating nomenclature and even some assistance. For example, people may forget that rejecting one option always means accepting another (if only the inaction option), which might prove even less attractive were it examined in detail. Conversely, accepting any option means foregoing others (e.g., because there are not enough resources to go around); the opportunity costs of doing without the net benefits that would be gained by adopting the rejected options need to be considered when evaluating the attractiveness of the one that is adopted (Vaupel & Graham, 1981). The Bayesian framework forces one to consider at least two options (and offers a label for failure to do so).

As mentioned earlier, the critical value provides a threshold for interpreting one's posterior odds: if they are above it, act as though H were true; if they are below, act as though $\bar{H}$ were true, considering the consequences of being right and wrong in either case. Thus, the critical value is the Bayesian way to relate uncertain knowledge about what may or may not be true to concrete actions, which either are or are not taken. As such it offers a criterion for judging whether the action taken is in one's own best interest. It may also provide something of a trap if it encourages people to confuse "acting as though H is true" with "believing that H is true." Such confusion may afflict decision makers who, having adopted a course of action, fail to attend to signals indicating that their best guess about H may have been wrong--and require revision. It may also characterize scientists who forget the uncertainties that they themselves acknowledge before offering a best-guess interpretation of some experimental results.

Complications

In the presentation above, the interpretation of a datum is complete once one has assessed the two conditional probabilities comprising the likelihood ratio. Such an appraisal assumes that the datum is taken at face value. More sophisticated Bayesian models are available for situations in which that assumption seems dubious--and the interpretation of data depends upon contextual factors. Two such elaborations that seem most pertinent to judgmental research deal with source credibility and conditional independence. These complications also point to some limits to the usefulness of the Bayesian scheme.

Source Credibility

Every datum comes from some source. Knowing that source may, in principle, have quite diverse effects on the datum's interpretation. In common parlance, one speaks about sources that have unusual or limited credibility, as well as about ones that may attempt to mislead or may have been misled themselves. On the basis of detailed modeling of the informational properties of evidentiary situations that may arise in the courtroom, Schum (e.g., 1980) has shown how source credibility information may reduce, enhance, or even reverse the diagnostic impact of a particular datum. The subtlety of Schum's models suggest both the difficulty of applyling Bayesian inference properly and the pitfalls awaiting those trying to rely on intuition. Without understanding the impact of source credibility information or asserting that a datum is taken at face value, it is hard to know how it is or should be interpreted.

A common task in judgment research requires participants to decide whether a target individual belongs to Category A or Category B on the basis of a brief description and some base-rate information. These descriptions vary along dimensions such as the internal consistency of the information they contain and the ratio of relevant to irrelevant information. Typically, investigators have considered only the informational content of these messages when analyzing the impact that these variations should and do have on behavior. In principle, however, each shift in content could (inadvertently) signal a different level of credibility. If subjects are sensitive to these signals, they may choose to respond in ways that are at odds with those dictated by the informational content--and be justified in doing so.

For example, Manis, Dovalina, Cardoze, and Avis (1980), as well as Ginosar and Trope (1980), varied the consistency of the information in the description. With consistent profiles, all information that pointed toward any category pointed toward the same category; with inconsistent profiles, such diagnostic bits pointed in both directions. Subjects relied less on inconsistent information. This might reflect sensitivity to its apparently lower diagnosticity; or it might reflect doubt about its overall credibility. If one doubts that inconsistent people exist (Cooper, 1981), one may discount sources that have produced descriptions showing inconsistency. Both responses could be justified normatively and would lead to similar judgments. However, they suggest different psychological processes. The latter interpretation would mean that this is a situation in which people do understand the need to regress predictions based on unreliable information (Kahneman & Tversky, 1973).

In situations where the content of a message provides a valid cue regarding its validity, failure to consider that possibility may mean mistaking sensitivity for bias. For example, in an experimental study of manuscript reviewing, Mahoney (1977) berated his scientist subjects for being more hospitable toward a fictitious study when its reported result confirmed the dominant hypothesis in their field than when it disconfirmed it. This differential receptiveness could reflect the stodginess and prejudices of normal science (Kuhn, 1962) which refuses to relinquish its pet beliefs. However, it could also reflect a belief that investigators who report disconfirming results tend to use inferior research methods (e.g., small samples, leading to more spurious results), or to commit common mistakes in experimental design, or simply to be charlatans. Mahoney himself might set a double standard when told that a study did or did not confirm the existence of telekinesis.

These reinterpretations are, of course, entirely speculative. In order to discipline them by fact, one needs to discover (a) how subjects are structuring the problem (e.g., are they worried at all about source credibility) and (b) how they appraise the different components of the inferential model they are using. For either describing or evaluating behavior, one must establish both what people believe and what they try to do with those beliefs.

Conditional Independence

The most general way of thinking about contextual effects is as an interaction between the meaning of two or more data. That is, one datum, D_i, creates a context that affects the interpretation of another datum, D_j. In Bayesian terms, such interactions are said to reflect conditional non-independence, because the conditional probability $P(D_j|H)$ is not equal to $P(D_j|H,D_i)$. As a result, one cannot compute the cumulative impact of a set of data simply by multiplying their respective likelihood ratios.

Source credibility problems (described above) may be viewed as a special case of conditional non-independence; information about the source affects interpretation of the message. Conversely, the message may affect one's view of the source (increasing or decreasing its apparent credibility). Conditional non-independence is also the grounds for configural judgment, the focus of many studies of clinical diagnosis (Goldberg, 1968). For the configural judge, the meaning of a particular cue depends upon the status of others (e.g., "that tone of voice suggests 'not suicidal' to me unless I know that it was spoken at midday"). The research record shows that, although clinicians claim to interpret cues configurally, firm evidence of configural judgment is hard to find. This may reflect the insensitivity of the research tool, the inaccuracy of the clinicians' introspections about their own judgmental processes, or their failure to use their configural strategies consistently (Dawes, 1979; Slovic & Lichtenstein, 1971).

In studies of clinical diagnosis, behavior is typically described by linear regression equations applied to large sets of structurally similar judgment tasks. Configural relations are represented in those equations by interaction terms. In the Bayesian model, conditional non-independence is treated by assessing joint conditional probabilities which consider interrelated data simultaneously. How these interrelationships are modeled is, of course, a matter of individual opinion. For that opinion to be trustworthy, it needs to be the product of careful application. Considering the subtleties of Schum's analyses of source credibility problems, the opportunities for error would seem to be many. Paradoxically, this very complexity may also mean that sets of interrelated data may defy explicit modeling, leaving them the province of judgment. One has to take a best guess at what the data mean, knowing that their mutual implications have not been understood.

Limits

Just as there are practical limits to the informational complexities that can be treated adequately within the Bayesian framework, so are there value considerations that are best recognized and left alone. In their lives (and in our experiments) people are not just acquiring knowledge for the sake of optimizing their actions. There are some pursuits that are not sensibly accommodated in the Bayesian framework and which can lead to deliberately non-Bayesian behavior. People may act suboptimally in the short run when they are pursuing such long-run goals as "maintaining social relations (e.g., preserving and cultivating information sources), gaining and sustaining recognition (e.g., exuding confidence where accountability is low), and being accepted (e.g., by passing up smart solutions that make one appear out of step)" (Fischhoff, 1981, p. 902). Thus, people may ask non-diagnostic questions in order to keep the conversation going; and they may pass up diagnostic ones because asking them seems untoward.

With sufficient ingenuity, such behavior could probably be translated into (Bayesian) decision theory terms so as to show that it is not just purposeful, but also rational--in the sense of having the highest expected value of any course of action. Thus, for example, the asking of a particular question could be treated as an act that has consequences, such as the cost of asking it (and interpreting the answer) and the penalty of being censured for impertinence. Practically speaking, it would be very hard to anticipate and analyze such a complex of considerations. Theoretically speaking, the attempt to do so may have a very ad hoc character, groping for causes that might be shaping (or deflecting) people's hypothesis testing. To be useful, these interpretations must walk a tightrope between giving people too little credit and giving them too much. At the former extreme, any behavior that has no ready Bayesian expression reflects cognitive incompetence. At the other extreme, people do whatever is right for them and the observer's task is to determine what it is that they were trying to (and managed to) optimize (Cohen, 1981 and Commentary; Hogarth, 1981).

Implications and Reinterpretations

On a prescriptive level, the Bayesian approach provides a general model of how people should be making sequential inferences. Using the model descriptively requires a choice between two strategies. One is to assume that people are intuitive Bayesians and explore how they use the model (e.g., what diagnostic value do they attribute to a particular datum). Or, one can assume that they do not use the model, however, their judgments can be described in terms of systematic departures from it (e.g., a number of apparently diverse phenomena were characterized as special cases of failing to consider the denominator in the likelihood ratio). In the course of explicating the Bayesian model, we have used both strategies, at times applying them to the same observed behavior. In some of these cases, that behavior could be viewed either as Bayesian or non-Bayesian, depending upon what one assumes about what people believe and what problem they are solving.

The present section exploits this framework to analyze the behavior that has been observed in a number of studies. In most cases, it casts a somewhat different light on what subjects in those studies were (and should have been) doing than was advanced by the original authors. In some cases, this reinterpretation shows commonalities in effects that had appeared to be distinct. In others, it reveals the differences in tasks that had gone under the same label. In particular, it shows how the term "confirmatory bias" has been applied to a variety of phenomena that may be described more succinctly in terms of the Bayesian model.

Nisbett, Zukier & Lemley (1981)

The stimuli in this series of studies were thumbnail descriptions of fictional individuals. The hypotheses were of the form "the individual belongs to category A" (e.g., is a child abuser) or "the individual belongs to category B" (e.g., is not a child abuser). Stimuli and hypotheses were designed to test the authors' attempted integration of Kahneman and Tversky's (1972) work on representativeness with Tversky's (1977) work on similarity judgments. According to the former, the judged probability of belonging to a category depends upon the judged similarity between the salient features of the description and of the category stereotype. According to the latter, judged similarity should increase with the number of salient features common to the individual's description and the category stereotype; it should decrease with the number of features unique to each.

Nisbett et al.'s descriptions consisted of one or two diagnostic features, pointing toward one of the two possible prediction categories, and a varying number of non-diagnostic features, pointing toward neither category. The authors found that as the number of non-diagnostic features increased, the impact of the diagnostic features decreased. Thus, the non-diagnostic data "dilute" the effect of the diagnostic data, leading to less confident predictions. As the authors note, this dilution could only be justified normatively if subjects perceived (what would be called here) conditional non-independence among the descriptors; that is, if the non-diagnostic information somehow mitigated the impact of the diagnostic information (e.g., by reminding subjects how complex people are, hence how unreasonable it is to make confident predictions on the basis of a single feature). Establishing the normative status of the observed dilution would require measuring the judged configurality of the entire data set (a difficult task which was not undertaken).

Although the logical status of these results is quite clear, the role of similarity judgments in them is not. If subjects rely on representativeness, the judged probability of an individual belonging to a category will depend upon apparent individual-category similarity. As described by Tversky (1977), similarity judgments do, indeed, concern such a match between _an_ individual and _a_ category. Nisbett et al. extend these notions to cover multiple possible categories by assuming that people form a subjective likelihood ratio whose numerator and denominator are derived separately by representativeness from judged individual-category similarity.

Nisbett et al. designed their stimuli so that non-diagnostic data belonged to neither category stereotype. As a result, addition of a non-diagnostic datum to a diagnostic datum should produce a description that is less similar to both category stereotypes than was the single datum description. The judged probability of getting such a description given that the individual belongs to each category should be correspondingly lower. If both conditional probabilities are reduced by equivalent amounts, then there should be no reduction in diagnosticity, hence no dilution effect. Given that the addition of non-diagnostic information did reduce the extremity of judgments, one must forfeit some component of the above account. One way of doing so is not to assume that subjects analyze evidence by forming subjective likelihood ratios. Rather, here, as elsewhere, subjects have ignored the denominator when considering diagnosticity. By doing so, they would fail to notice that the non-diagnostic information also reduced the match between the description and the competing category.

Such neglect of the denominator would have two additional implications for the interpretation of Nisbett et al.'s study. One is that it is unclear that pretest subjects were judging "diagnosticity," in the conventional sense, when they appraised "how helpful the information would be for prediction" (p. 255). The second is that reliance on representativeness in the manner suggested by Nisbett et al. might prove to be a modest aid to inference. People asked to assess the likelihood that an individual is from Category A (H) or Category B ($\bar{H}$) on the basis of a brief personality description (D) might be encouraged to rely on representativeness. Asking whether D is more representative of A or B would mean making a judgment that has at least the elements of the likelihood ratio (D, H, $\bar{H}$), if not necessarily in the proper relationship to each other. They might further be cautioned against the main problem that seems to arise with reliance on this strategy, the tendency to neglect the prior odds whenever one can detect any degree of differential representativeness (Bar-Hillel & Fischhoff, 1981; Ginosar & Trope, 1981).

Snyder & Swann (1978)

In this series of studies, subjects selected questions that "would provide them with the information to best test the hypothesis" (p. 1204) that a target individual whom they were about to interview was an introvert (or an extravert). In the subjects' choice of questions, the authors claimed to have demonstrated a new bias, namely "an erroneous tendency to search for evidence that would tend to confirm the hypothesis under scrutiny" (p. 1203).

From a Bayesian perspective, however, it is unclear how the choice of question could be biased toward confirmation of the focal hypothesis. Mathematically, it is not possible to ask questions all of whose possible answers will support a particular hypothesis. Thus, there is no confirmatory bias in the sense of sampling data that will inevitably favor a particular hypothesis (as opposed to sampling data that will favor its competitors). Snyder and Swann apparently felt that their subjects are asking questions whose answers they can anticipate and will interpret as supporting the focal hypothesis. If the answer, D_i, to a question is predictable, then $P(D_i) = 1 = P(D_i|H)/P(D_i|\bar{H})$. That is, the question is non-diagnostic--and the search process inefficient. If D_i is taken as evidence supporting H, then the problem lies with the interpretation, perhaps reflecting neglect of the denominator in the likelihood.

A less extreme form of this exploitation of predictability might be seen in a scientist who repeatedly replicates the same experiment which typically produces the same observation, D, which is more likely given H than given $\bar{H}$. Each observation of D would not, however, provide an independent and equal contribution to affirming H. An observation is informative only to the extent that it is unpredictable, hence capable of affirming or disaffirming the hypothesis. With each replication of the experiment, P(D) increases, independent of the truth of H. There is a corresponding decrease in the disparity between $P(D|H)$ and $P(D|\bar{H})$ that is necessary for a diagnostic likelihood, until after many replications P(D) approaches 1. An analogous way to look at this problem is to view the likelihood ratio as a measure of association defined on a 2 x 2 table whose rows and columns are $(D, \bar{D})$ and $(H, \bar{H})$, respectively. As the marginals of one of the variables (here D) become more extreme, the value of the measure of association tends to decrease.

Perhaps the only way to bias the search toward confirmation would be to ask questions for which $P(D|H)$ and $P(D|\bar{H})$ were expected to be high, knowing that one would subsequently ignore $P(D|\bar{H})$. Such a biased-search-in-the-service-of-biased interpretation could, of course, only be exhibited successfully when P(D) is high (i.e., it is likely to be observed whether or not H is true). If subjects tried it when P(D) was low, then it would represent a disconfirmation bias: subjects would only look at $P(D|H)$, which would be low, thereby reducing their confidence in H.

In Snyder and Swann's experiment, subjects asked to test the hypothesis that an individual was an extravert were counted as biased if they chose "confirmatory" questions such as "what would you do if you wanted to liven up a party?" rather than "disconfirmatory" questions such as "what factors make it hard for you to really open up to people?" or "neutral" questions such as "what are your career goals?" These non-neutral questions have two noteworthy properties. One is that they are non-diagnostic, insofar as they would elicit similar responses from both introverts and extraverts (and unless there are particularly introverted ways to liven up parties). Such questions are inefficient, but their selection constitutes no bias toward confirmation. As there were only 5 neutral questions in the set of 26 possibilities, it is not surprising that subjects chose many non-neutral ones in their set of 12 test questions. Their preference for confirmatory over disconfirmatory non-neutral questions may have reflected some crude compatibility with the task hypothesis.

A second peculiar property of the non-neutral questions is that they are phrased in a conditional way that assumes the category membership of the recipient. For example, it would be awkward or untoward to ask an introvert "what would you do if you wanted to liven up a party?" (Trope & Basok, in press). Perhaps the only way one might get any useful information with such a question would be if respondents were to reject its premises where appropriate (e.g., "What do you mean? I never want to liven up parties!").[10] The fact that subjects did not ask all the neutral questions before proceeding to these uncomfortable non-neutral ones may reflect the undiagnostic nature of some of the neutral questions (e.g., "To what kinds of charities do you like to contribute?").

Wason (1960, 1968)

Card task. For some twenty years, Wason and colleagues (see Evans, in press) have been studying how people test what might be called *formal* or *logical* hypotheses. These are categorical statements, such as "All As are Bs," which may be disconfirmed by a single counter-example (e.g., an A that is not a B). In one popular experiment, subjects are told that each of four cards has either A or $\bar{A}$ on one side and either B or $\bar{B}$ on the other side. They are then shown the four cards, presented so that the exposed sides show symbols that are, respectively, A, $\bar{A}$, B, and $\bar{B}$. Subjects' task is to choose the cards they would turn over in order to test the hypothesis that "all As are Bs." Even though turning over either the A or the $\bar{B}$ could falsify the hypothesis, most people "waste" one of their choices on the B card, which is necessarily inconclusive. From a Bayesian perspective, this bias is an example of both asking a non-diagnostic question (B) and failing to ask a potentially falsifying question ($\bar{B}$).

A task such as Wason's constitutes a special case of Bayesian inference in at least two senses. One is that with logical hypotheses, it is possible to observe data that unambiguously falsify or verify an hypothesis (i.e., that have a likelihood ratio of zero or infinity). With empirical hypotheses, regarding real-life events, such certitude is rare or impossible. Indeed, one might speculate that subjects fail to ask potentially falsifying questions because they are unaccustomed to testing logical hypotheses (and to having falsification be a possibility). Second, the artificiality of the problem gives subjects no basis for assigning most of the probabilities involved. Many different priors could be defended on the basis of the imputed likelihood that the experiment would focus attention on a true hypothesis. Except for data that would disconfirm H or $\bar{H}$, it is hard to defend any value of the likelihood ratio, insofar as any value other than 0 and infinity requires some arbitrary assumptions about the correlation between A-ness and B-ness in the artificial universe that the experimenter has created.

Triads task. In another task devised by Wason, subjects are told that the numbers 2, 4, and 6 conform to a simple rule that the experimenter has in mind; their task is to guess that rule. As an aid, they may ask whether additional number triads of their own choosing conform to the rule. In such a logical task, the validity of a hypothesized rule cannot be proved. Even if a rule fits all conforming triads, there is no guarantee that some other rule might not also fit, nor that some future triad might not violate it. It is only subjects' suppositions about the sorts of rules that the experimenter might use and the negligible penalty for guessing wrong that would enable strictly Bayesian subjects to stop gathering information (i.e., stop proposing triads) and guess that a particular hypothesis is correct.

The experimenter's (unannounced) rule is "numbers increasing in value." The first rule that comes to most people's minds is apparently "sequential even numbers." The modal first triad is something like 8, 10, 12. Called a sign of "verification bias" by Wason, this response pattern may be explained by several different accounts.

The first account argues that subjects believe that being told that 8, 10, 12 fits the rule will prove that "sequential even numbers" is the correct hypothesis (whereas being told that it does not fit will prove that "sequential even numbers" is not the correct hypothesis). Hence, they propose 8, 10, 12 expecting to get a definitive answer, either way. This flaw in logical inference could be given a special name, such as mistaking affirmation for confirmation. However, it is most parsimoniously regarded as yet another instance of ignoring alternative hypotheses when evaluating evidence. For the datum "yes, it conforms" and the hypothesis "the rule is sequential even numbers," $P(D|H) = 1$. Hence, that answer cannot reduce one's confidence in the truth of H. The evidence becomes, of course, less conclusive when one realizes that many alternative hypotheses (including the experimenter's own) would evoke the same response and have the same associated conditional probability.

A second account is a Bayesian interpretation of Wason's own. It holds that subjects seek evidence that will enable them to compare their hypothesis with its complement (all other rules).[11] The question they ask (8, 10, 12) does, in fact, produce information that is diagnostic for this comparison. Where subjects fail is in not asking one of those even more diagnostic questions which could falsify their hypothesis.

The third account has subjects comparing their favored hypothesis with another specific hypothesis, acting for the moment as though those two exhausted the universe of possibilities. Although they might be faulted for making such a knowingly false assumption, subjects might still consider this simplification a useful fiction. A more "proper" strategy is to compare each hypothesis with its complement and then, should falsifying evidence be found, start afresh with a new hypothesis. However, in a situation where any one of an infinite number of hypotheses could be the true one, people may prefer to compare pairs of hypotheses sequentially; at each round, the less likely hypothesis is discarded and the more likely one is then compared with the next contender. Such a comparative strategy would be similar in spirit to sophisticated falsificationism (Lakatos, 1970), whereby one holds onto the hypothesis that seems most correct until a better candidate comes along, even to the point of holding onto an hypothesis for which inconsistent data are known to exist. The "proper" strategy resembles naive falsificationism (Popper, 1972), whereby one focuses on a single hypothesis, doing everything possible to disprove it without regard to what might take its place.

Subjects using this comparative strategy could still be faulted for poor triad selection. For any pair of hypotheses, they should ask about triads for which an affirmative answer will falsify one hypothesis and a negative answer will falsify the other. Thus, 8, 10, 12 is a poor triad if the two hypotheses are "sequential even numbers" and "numbers increasing in value." The experimenter's response would be "yes" if either were the correct hypothesis. On the other hand, it would be a fine choice if the two hypotheses were "sequential even numbers" and "numbers less than 7." Without eliciting subjects' hypotheses, it is hard to tell whether to fault them for using the comparative strategy, or for using it inefficiently (by choosing non-diagnostic triads).

Conclusion

Bayesian inference was originally developed as a prescriptive model. Its advocates believe that when engaging in a sequential inference task it is useful to identify each element in the Bayesian model with the corresponding elements in one's thinking. Such identification ensures that all necessary elements have been considered and that they have been put in the proper relationship to one another. Although the model can guide people in structuring their hypothesis testing, it cannot help them in formulating their hypotheses, assessing component probabilities or setting the critical ratio. Performing these operations requires a substantive understanding of the problem at hand.

Like other prescriptive models, the Bayesian scheme assumes that people could follow its dictates if they tried and were given some feasible level of assistance. However, it makes no statement regarding how people actually do make decisions. The descriptive potential of the Bayesian model lies in the nomenclature that it provides for the primitives and processes of people's intuitive inference. The present essay attempts to exploit this potential by identifying the kinds of systematic deviations from the Bayesian model that could, in principle, be observed. Illustrative examples of most of these biases are found in the research literature, some collected within a Bayesian framework, some not.

Once developed, this descriptive model is then applied to a variety of existing studies. Although the results of this (re)interpretation are typically specific to particular studies, a number of general conclusions seem to emerge:

(a) In some situations, a number of apparently diverse effects have proven to be special cases of a particular judgmental bias. The most powerful of these "metabiases" is the tendency to ignore $P(D|\bar{H})$ when evaluating evidence.

(b) In some situations, a variety of different phenomena have been confused under a common title. "Confirmation bias," in particular, has proven to be a catch-all phrase incorporating biases in both information search and interpretation. Because of its excess and conflicting meanings, the term might best be retired.

(c) In all situations, a careful appraisal of how judges have interpreted the tasks posed to them is needed before making any assertions regarding how, if at all, their judgment is biased. In making such speculations about how people may have construed particular tasks, it is important to strike a balance between exaggerating the extent to which we or they are all-knowing. Neither our understanding of people nor our ability to help them is served by uncritically assuming either that there is no way for them to justify behavior that seems suboptimal to us or that there is a hidden method to any apparent madness that people exhibit.

Reference Note

1. Moskowitz, H., & Sarin, R. K. Improving conditional probability assessments for long range forecasting and decision making. Paper No. 734. Institute for Research in the Behavioral, Economic and Managment Sciences. Purdue University, 1980.

References

Bar-Hillel, M. The role of sample size in sample evaluation. Organizational Behavior and Human Performance, 1979, 24, 245-257.

Bar-Hillel, M. The base-rate fallacy in probability judgments. Acta Psychologica, 1980, 44, 211-213.

Bar-Hillel, M., & Fischhoff, B. When do base rates affect predictions? Journal of Personality and Social Psychology, 1981, 41, 671-680.

Beyth-Marom, R. How probable is probable? Numerical translation of verbal probability expressions. Journal of Forecasting, in press.

Borgida, E., & Brekke, N. The base-rate fallacy in attribution and prediction. In J. H. Harvey, W. J. Ickes, and R. F. Kidd (Eds.), New directions in attribution research (Vol. 3). Hillsdale, N.J.: Lawrence Erlbaum, 1981.

Brown, R. V., Kahr, A. S., & Peterson, C. Decision analysis for the manager. New York: Holt, Rinehart & Winston, 1974.

Business Week. Friedman denies linkage policy. March 18, 1979.

Cohen, J. Can human irrationality be experimentally demonstrated? The Behavioral and Brain Sciencs, 1981, 4, 317-370.

Cooper, W. H. Ubiquitous halo. Psychological Bulletin, 1981, 90, 218-244.

Dawes, R. M. The robust beauty of improper linear models in decision making. American Psychologist, 1979, 34, 571-582.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. Pseudodiagnosticity. Acta Psychologica, 1979, 43, 111-121.

Doyle, A. C. The memoirs of Sherlock Holmes. London: John Murray and Jonathan Cape, 1974 (originally published 1893).

Eddy, D. M. Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Edwards, W. Dynamic decision theory and probabilistic information processing. Human Factors, 1962, 4, 59-73.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review*, 1963, *70*, 193-242.

Evans, J. St. B. T. *The psychology of deductive reasoning*. London: Routledge & Kegan Paul, in press.

Fischer, D. H. *Historian's fallacies*. New York: Harper & Row, 1970.

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental psychology: Human Perception and Performance*, 1975, *1*, 288-299.

Fischhoff, B. Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 349-358.

Fischhoff, B. For those condemned to study the past: Reflections on historical judgment. In R. A. Shweder & D. W. Fiske (Eds.), *New directions for methodology of behavior science: Fallible judgment in behavioral research*. San Francisco: Jossey-Bass 1980.

Fischhoff, B. Inferential interference. Review of *Human inference: Strategies and shortcomings of social judgment*, by R. Nisbett and L. Ross. *Contemporary Psychology*, 1981, *26*, 901-903.

Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press, 1982.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 330-344.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing what you want: Measuring labile values. In T. Wallsten (Ed.), *Cognitive processes in choice and decision behavior*. Hillsdale, N.J.: Erlbaum, 1980.

Forer, B. The fallacy of personal validation: A classroom demonstration of gullibility. Journal of Abnormal and Social Psychology, 1949, 44, 118-123.

Freedman, L. The evolution of nuclear strategy. London: MacMillan, 1981.

Ginosar, Z., & Trope, Y. The effects of base rates and individuating information on judgments about another person. Journal of Experimental Social Psychology, 1980, 16, 228-242.

Goldberg, L. R. Simple models or simple processes? Some research in clinical judgment. American Psychologist, 1968, 23, 486-496.

Green, A. E., & Bourne, A. J. Reliability technology. New York: Wiley Interscience, 1982.

Hogarth, R. M. Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 1981, 90, 191-217.

Hyman, R. Cold reading. Zetetic (The Skeptical Inquirer), 1977, 1(2), 18-37.

Kahneman, D., & Tversky, A. Subjective probability: A judgment of representativeness. Cognitive Psychology, 1972, 3, 430-454.

Kahneman, D., & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-251.

Kahneman, D., & Tversky, A. Intuitive prdictions. Biases and corrective procedures. TIMS Studies in Management Science, 1979, 12, 313-327.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Kish, L. Survey sampling. New York: Wiley, 1965.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 1980, 6, 107-118.

Kuhn, T. Structure of scientific revolutions. Princeton: Princeton University Press, 1962.

Kyburt, H. E., & Smokler, H. E. (Eds.), Studies in subjective probability. Huntington, N.Y.: Robert E. Krieger, 1980.

Lakatos, I. Falsification and scientific research programs. In I. Lakatos and A. Musgrave (Eds.), Criticism and the growth of scientific knowledge. Cambridge: Cambridge University Press, 1970.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: State of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4, 551-578.

Lindley, D. V., Tversky, A., & Brown, R. V. On the reconciliation of probability assessments. Journal of the Royal Statistical Society, Series A, 1979, 142, Pt. 2, 146-180.

Lyon, D., & Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. Acta Psychologica, 1976, 40, 287-298.

Mahoney, M. J. Publication prejudices. Cognitive Therapy and Research, 1977, 1, 161-175.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. Base rates can affect individual predictions. Journal of Personality and Social Psychology, 1980, 38, 231-240.

Mehle, T., Gettys, C. V., Manning, C., Baca, S., & Fisher, S. The availability explanation of excessive plausibility assessments. Acta Psychologica, 1981, 49, 127-140.

Murphy, A. H. Scalar and vector partitions of the probability score: Two-state situation. Journal of Applied Meteorology, 1972, 11, 273-282.

Nisbett, R., & Ross, L. Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Nisbett, R. E., Zukier, H., & Lemley, R. E. The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. Cognitive Psychology, 1981, 13, 248-277.

O'Leary, M. K., Coplin, W. D., Shapiro, H. B., & Dean, D. The quest for relevance. International Studies Quarterly, 1974, 18, 211-237.

Phillips, L. D. Bayesian statistics for social scientists. London: Nelson, 1973.

Pitz, G. F., Downing, L., & Reinhold, H. Sequential effects in the revision of subjective probabilities. Canadian Journal of Psychology, 1967, 21, 381-393.

Popper, K. R. Objective knowledge. Oxford: Clarendon, 1972.

Raiffa, H. Decision analysis. Reading, Mass.: Addison-Wesley, 1968.

Reyna, V. F. The language of possibility and probability: Effects of negation on meaning. Memory and Cognition, 1981, 9, 642-650.

Schlaifer, R. Analysis of decisions under uncertainty. New York: McGraw-Hill, 1969.

Schum, D. Current developments in research on cascaded inference processes. In T. Wallsten (Ed.), Cognitive processes in choice and decision behavior. Hllsdale, N.J.: Erlbaum, 1980.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Snyder, M., & Swann, W. B. Hypothesis testing process in social interaction. Journal of Personality and Social Psychology, 1978, 36, 1202-1212.

Stech, F. J. Intelligence, operations and intentions. Military Intelligence, 1980, 37-43.

Sue, S., Smith, R. E., & Caldwell, C. Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. Journal of Applied Social Psychology, 1973, 3, 345-353.

Trope, Y., & Basok, M. Confirmatory and diagnosing strategies in social-information gathering. Journal of Personality and Social Psychology, in press.

Troutman, C. M., & Shanteau, J. Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, 1977, *19*, 43-55.

Tversky, A. Features of similarity. *Psychological Review*, 1977, *84*, 327-352.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 1974, *185*, 1124-1131.

Tversky, A., & Kahneman, D. Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology*. Hillsdale, N.J.: Erlbaum, 1980.

Tversky, A., & Kahneman, D. The framing of decisions and the rationality of choice. *Science*, 1981, *211*, 453-458.

U.S. Nuclear Regulatory Commission. *Reactor safety study: An assessment of accident risks in U.S. commercial nuclear power plants*. WASH-1400, NUREG-75/014. Washington D.C.: The Commission, 1975.

U.S. Nuclear Regulatory Commission. *Risk assessment review group report to the U.S. Nuclear Regulatory Commission*. NUREG/CR-0400. Washington, D.C.: The Commission, 1978.

U.S. Nuclear Regulatory Commission. *Fault tree handbook*. Washington, D.C.: The Commission, 1980.

Vaupel, J., & Graham, J. Eggs in your bier. *The Public Interest*, 1981.

Wason, P. C. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 1960, *12*, 129-140.

Wason, P. C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 1968, *23*, 273-281.

Acknowledgements

Partial support for this research came from the UK Medical Research Council and from the U.S. Office of Naval Research under Contract N00014-80-C-0150 to Perceptronics, Inc. Correspondence may be addressed to Baruch Fischhoff, MRC/APU, 15 Chaucer Road, Cambridge CB2 2EF, England, or to Ruth Beyth-Marom, Decision Research, A Branch of Perceptronics, 1201 Oak St., Eugene, Oregon 97401 U.S.A.

Footnotes

[1] The demand of coherence is what differentiates DeMorgan's "pure subjective" interpretation of probability--as whatever people actually believe--from DeFinetti's personalistic interpretation of probability as rational belief.

[2] A topical example may be found in the runaway inflation that has followed the linking of incomes and loans to the cost-of-living index in several countries. Although his economic theories predicted the opposite result, Milton Friedman has denied that this unhappy experience constitutes evidence against his theories because the countries involved did not implement the indexation exactly as he prescribed. Even more troubling for the status of his theories about the economy is that further thought (perhaps stimulated by this irrelevant experience) has led him to conclude that his earlier derivation was wrong and that, in fact, indexation encourages inflation (<u>Business Week</u>, 1979).

[3] In the case of multiple hypotheses, the posterior probability $P(H_i|D) = (P(D|H_i)P(H_i)/\sum_i P(D|H_i)P(H_i)$. A diagnostic datum is then one for which $P(D|H_i) \neq \sum_i P(D|H_i)P(H_i)$.

[4] It may be worthwhile to mention in passing one difficulty that the Bayesian approach avoids--vagueness in expressing beliefs. It has been found that people disagree considerably about the interpretation of verbal expressions of likelihood (e.g., "probably true"). Moreover, the same individual may use a term differently in different contexts (Beyth-Marom, in press; Reyna, 1981). The bayesian approach requires explicitness.

[5] It may be a useful aside to note that a subjectivist would deny the objectivity of even a probability such as that assigned to the flip of a coin following the observation of a long series of flips. The decision to aggregate those flips into an overall "probability estimate" and then to apply that estimate to the next flip requires a (subjective) assessment that those trials were identical in all relevant aspects. Consensus on their identicalness does not guarantee its truth.

6. One might argue that the collection of diverse pieces of existing evidence should not be considered the assessment of a base rate, a term that should be reserved for aggregating statistical data (e.g., 70% of previous cases have supported H). Following that argument would lead to the Laplacian assumption that all hypotheses are equally likely a priori, except in the presence of statistical data to the contrary. In that case, the problems discussed in the text could be relegated to the section on assessing likelihood ratios (associated with additional pieces of data). Of course, one might also argue that the statistical data themselves were separate pieces until they were aggregated--meaning that their aggregation required the use of likelihood ratios. At this extreme, a priori odds would always be equal to 1.

7. The undue influence of prior information in this context is in sharp contrast to the neglect of prior information observed with the base-rate fallacy. The crucial difference between the two cases is that in the former the prior odds are actually posterior odds that one has arrived at by actively weighing previous evidence, whereas in the latter the prior is but a statistical summary, reporting what someone else has typically observed.

8. The contrast between examples (b) and (c) shows that the same bias, neglecting the likelihood ratios denominator, can produce substantively different behavior. In (b), it led to irrelevant data reducing faith in an hypothesis; in (c), it led to increased faith following receipt of a datum that should have reduced it.

9. Deliberately unrepresentative sampling would seem to be another possible bias to be included in this section. However, we will argue below (in the discussion of Snyder & Swann, 1978), that such biases are better conceptualized as problems of data interpretation than problems of sampling.

10. Only in Experiment 2 did Snyder and Swann's subjects actually conduct the promised interview with the target person. Other subjects who listened to tapes of these interviews judged

respondents to predominantly extravert questions to be more extraverted than respondents to predominantly introvert questions. One might concur with Snyder and Swann's speculation that the biased question sample evoked a biased sample of reported behavior. Or, one may speculate that the form of the interviewees' responses reflected the conditioning presumptions of the questions they were asked. That is, it may be relatively easy to tell when people are answering questions that "would typically be asked of people *already known* to be extraverts" (p. 1204; emphasis in original).

11. It would not be proper to compute a likelihood ratio comparing the impact of the evidence on the subjects' hypothesis and the experimenter's hypothesis, because they are not mutually exclusive.

OFFICE OF NAVAL RESEARCH

TECHNICAL REPORTS DISTRIBUTION LIST

CDR Paul R. Chatelier
Office of the Deputy Under Secretary
of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Engineering Psychology Programs
Code 422
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217 (5 cys)

Manpower, Personnel & Training
Programs
Code 270
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Operations Research Programs
Code 411-OR
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics & Probability Program
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Systems Program
Code 411-IS
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

CDR K. Hull
Code 410B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Physiology & Neuro Biology Programs
Code 441
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Commanding Officer
ONR Eastern/Central Regional Office
ATTN: Dr. J. Lester
Bldg. 114, Section D
666 Summer Street
Boston, MA 02210

Commanding Officer
ONR Western Regional Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
American Embassy, Room A-407
APO San Francisco, CA 96503

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D.C. 20375

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis & Evaluation
Group
Naval Training Equipment Center
Code N-00T
Orlando, FL 32813

Dr. Albert Colella
Combat Control Systems
Naval Underwater Systems Center
Newport, RI 02940

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Dr. A.L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

Mr. Arnold Rubinstein
Naval Material Command
NAVMAT 0722 - Rm. 508
800 North Quincy Street
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 340F
Washington, D.C. 20361

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bathesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Department
Naval Medical Research Institute
Bethesda, MD 20014

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Dr. James McGrath
CINCLANT FLT HQS
Code 04E1
Norfolk, VA 23511

Navy Personnel Research &
Development Center
Planning & Appraisal Division
San Diego, CA 92152

Dr. Robert Blanchard
Navy Personnel Research &
Development Center
Command & Support Systems
San Diego, CA 92152

LCDR Stephen D. Harris
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA L8974

Dr. Julie Hopson
Human Factors Eningeering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

CDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, CA 93940

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations &
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks, AFB, TX 78235

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
ENGLAND

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsview, Ontario M3M 3B9
CANADA

Dr. A.D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
ENGLAND

Dr. Robert T. Hennessy
NAS-National Research Council
2101 Constitution Ave., N.W.
Washington, D.C. 20418

Dr. M.G. Samet
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91367

Dr. Robert Williges
Human Factors Laboratory
Virginia Polytechnic Institute
& State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Alphonse Chapanis
Department of Psychology
The Johns Hopkins University
Charles & 34th Streets
Baltimore, MD 21218

Dr. Ward Edwards
Director, Social Science Research
Institute
University of Southern California
Los Angeles, CA 90007

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, D.C. 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Christopher Wickens
University of Illinois
Department of Psychology
Urbana, IL 61801

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 cys)

Dr. Judith Daly
System Sciences Office
Defense Advanced Research Projects
Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goleta, CA 93017

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, CO 80309

Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, VA 22202

Dr. T. B. Sheridan
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Harry Snyder
Department of Industrial Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. W. S. Vaughan
Oceanautics, Inc.
422 6th Street
Annapolis, MD 21403

Dr. Richard W. Pew
Information Sciences Division
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Hillel Einhorn
University of Chicago
Graduate School of Business
1101 E. 58th Street
Chicago, IL 60637

Dr. John Payne
Duke University
Graduate School of Business
Administration
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Andrew P. Sage
University of Virginia
School of Engineering and
  Applied Science
Charlottesville, VA 22901

Dr. Leonard Adelman
Decisions and Designs, Inc.
8400 Westpark Drive, Suite 600
P.O. Box 907
McLean, VA 22101

Dr. Lola Lopes
Department of Psychology
University of Wisconsin
Madison, WI 53706

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Industrial Park
Burlington, MA 01803

Dr. Rex Brown
Decision Science Consortium
Suite 721
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Hillsdale, N.J.: Erlbaum, 1980.